ISSN 1503-299X

WORKING PAPER SERIES

No. 2/2022

The Importance of Tutors' Instructional Practices: Evidence from a Norwegian Field Experiment [1]

Hans Bonesrønning
Department of Economics
Norwegian University of Science and Technology

Jon Marius Vaag Iversen
NTNU Social Research

Department of Economics

Norwegian University of Science and Technology
N-7491 Trondheim, Norway
http://www.ntnu.edu/econ/working-papers

1) We are grateful to Henning Finseraas, Ines Hardoy, Ole Henning Nyhus, Vibeke Opheim, Kari Vea Salvanes, Astrid Marie Jorde Sandsør, and Pål Schøne for valuable contributions to the design and execution of the intervention. Comments from the Scientific Advisory Board appointed by the Research Council of Norway: Peter Fredriksson, Peter Blatchford, and Dorteh Bleses are highly appreciated. We are grateful to Ester Bøckmann for excellent research assistance. This research is part of the 1+1 Project, supported by the Norwegian Research Council under Grant 256217.

The Importance of Tutors' Instructional Practices: Evidence from a Norwegian Field Experiment[1]

Hans Bonesrønning

Norwegian University of Science and Technology

Jon Marius Vaag Iversen

NTNU Social Research

We use high quality black box data from a large Norwegian field experiment, where students in the early grades in turn were pulled out of their regular classes and offered mathematics instruction in small homogenous groups, to investigate how the tutors adapted their instruction to the size and the average performance level of the groups. Using within-tutor variation, we find that tutors tailored their instruction to the average pretest scores in the small groups and offered individualized instruction, especially to low achievers. We also find that the instructional practices varied substantially between the tutors, from teacher-directed to student-oriented practices. We show that the largest achievement gains were associated with a subgroup of tutors who spent much instructional time with evasive students in the low achievers' small groups ("inclusive individualization"). Finally, we show that the treatment effects were significantly larger in schools where the tutors practiced individualized and inclusive instruction for low achievers, compared to schools where the tutors had chosen student-centered practices and paid little attention to evasive students.

Keywords: *tutoring, tracking, instructional practices, mathematics*

[1] We are grateful to Henning Finseraas, Ines Hardoy, Ole Henning Nyhus, Vibeke Opheim, Kari Vea Salvanes, Astrid Marie Jorde Sandsør, and Pål Schøne for valuable contributions to the design and execution of the intervention. Comments from the Scientific Advisory Board appointed by the Research Council of Norway: Peter Fredriksson, Peter Blatchford, and Dorteh Bleses are highly appreciated. We are grateful to Ester Bøckmann for excellent research assistance. This research is part of the 1+1 Project, supported by the Norwegian Research Council under Grant 256217.

Student heterogeneity is a persistent challenge in all mass education systems. While some people worry that high achievers are being held back in heterogeneous and noisy classrooms, many more are concerned about students who struggle and leave school with poor basic skills. These worries are reflected in educational research. There is a burgeoning empirical literature on the effects of tutoring for struggling students and a thin, and much more immature empirical literature on the effects of tracking, probably reflecting the conventional wisdom that tracking favors high achievers, leaving low achievers worse off.

Much of the recent empirical research on tutoring is carried out as field experiments, thus generating credible results about treatment effects (see reviews by Dietrichson, Bøg, Figes, and Jørgensen, 2017, and Nickow, Oreopoulos, and Quan, 2020). Its usefulness is constrained by the lack of knowledge about the underlying mechanisms. Notably, much remains with respect to understanding teacher/tutor behaviors, and the effects of teacher/tutor behavior, under different circumstances. The purposes of this paper are to investigate whether tutors take advantage of small groups to provide relevant support to all students, and whether tutors take advantage of homogeneous groups to provide teaching that is tailored to their ability levels.

Several researchers (for example Guryan, Ludwig, Bhatt, Cook, Davis, Dodge, Farkas, Fryer, Mayer, Pollack, and Steinberg, 2021) motivate tutoring experiments by citing Bloom (1984), who states that teachers in regular classes tend to give the students in the upper third of the class the most attention and students in the lower third of the class the least attention and support. While tutoring one-on-one can do away with this problem, in tutoring two-on-one and in small groups the tutor must decide on the allocation of instructional time between the students. We are aware of no recent investigations of tutors' actual allocation of instructional time across students.

On a general level, researchers agree that tracking – in any form - leads to greater inequality in outcomes unless the benefits from tailored instruction dominate negative peer group effects for low achievers. Betts (2011), in his review of the empirical tracking literature, states that "In spite of many decades of research, what we do not know about the effects of tracking on outcomes greatly exceeds what we do know." In the present paper, we emphasize that effective tracking requires that the characteristics of effective instruction for students of different abilities is known and practiced. Recent contributions (Morgan, Farkas and Maczuga, 2015) have made progress on the first of these conditions. They separate between teacher-directed and student-centered instruction to show that students with mathematics difficulty (MD) benefit from teacher-directed instruction, while other student subgroups have equal positive effects for the two types of instruction. Related to the second condition - that effective instruction is practiced – Morgan, Farkas and Maczuga (op.cit) provide evidence that many teachers do not practice teacher-directed instruction for MD students.

Tracking faces additional challenges related to teacher behavior. Duflo, Dupas and Kremer (2011) show that many teachers provide less effort (i.e., are more absent) when faced by low achievers, and we know from tracking analyzes using non-experimental data that high-quality teachers prefer to teach high achievers.

The present paper contributes to the existing literatures by highlighting the instructional practices of tutors as observed in a large field experiment that combines tutoring and tracking. From a sample of 160 Norwegian elementary schools, 79 schools received -by a random draw- an extra (certified) teacher man-year to be used for small group instruction in mathematics for

students from 2nd to 4th grade (the students are from 7 to 9 years old). All students in the treated classes were pulled out of the regular class in turn and taught in groups of 4-6 students for two periods of 4-6 weeks per school year. The teachers (the regular teacher in cooperation with the tutor) were recommended to practice tracking – to form small groups that were homogenous with respect to a pre-treatment test score. Although tracking was not required, when asked in a survey, a large majority of the tutors agreed or strongly agreed that "the small groups are made up of students with almost the same skill level in mathematics". To keep unobservable teacher quality constant within schools, the same teacher-tutor should teach all small groups within a grade. The control schools were instructed to not change their practices.

Unlike several other recent experiments, the Norwegian experiment made use of tutors that were certified to teach mathematics. After being recruited to the experiment, the tutors were informed in project meetings about the characteristics of effective instruction. In addition, a Handbook was distributed to the regular teachers and the tutors at the start-up of the intervention. The principles of direct instruction - saying that sessions should be carefully organized step-by-step, with the steps building on each other – were part of it. Notably, presentation of new material should be followed by guided practice and feedback – the essential elements in mastery learning and individualization of instruction. Moreover, to encourage tailoring to low, medium, and high achievers the Handbook presented existing evidence about the characteristics of effective mathematics instruction for different subgroups of students. Also, it was pointed out that earlier successful interventions with tutoring for struggling students were characterized by tight collaboration between the regular teachers and the tutors.

Equipped with limited and quite general information about small group instruction for homogeneous small groups, we expect the tutors' professional skills to be crucial in turning treatment into student achievement. Thus, their abilities to *individualize and tailor* their instruction are at the center of the analyses presented here.

Individualization is about providing instruction based on the individual student's entering skill level, while tailoring refers to instruction that is adjusted to fit the average skill level of students in homogenous groups. Contrary to tailoring, individualization requires that the tutor makes decisions about the allocation of instructional time between the students in the group. We investigate whether the tutors take advantage of the small group size to reach out with individualized instruction to *all* students, hereinafter referred to as inclusive individualization or inclusive instruction. Since existing evidence indicates that teachers tend to give more attention to high achievers in regular classes, our basic hypothesis is that tutors' attention is skewed also in small groups that are homogenous with respect to a pretest score.

We capture the degree of tailoring by distinguishing between teacher-directed and student-centered practice. Morgan et al (op.cit.) cite Stein, Silbert and Carnine (2004) to define teacher-directed practices as "teachers helping students increase their procedural fluency in applying explicitly taught and repeatedly practiced sets of procedures to solve mathematics problems" (a specific example is routine practice and drill), and they cite Clements and Battista (1990) to define student-centered activities as providing "students with opportunities to be actively involved in the process of generating mathematical knowledge". We adhere to this definition, i.e., the core of the student-centered activities is that the students work together/ help each other to solve problems and to develop mathematical reasoning. In this paper tailoring is about

choosing the optimal mix of teacher-directed instruction and student-centered practices across student subgroups that differ with respect to initial mathematics skills.

We can address the instructional practices across and within the small groups because the tutors - in repeated surveys - are asked about the characteristics of the small group they are currently teaching, and the characteristics of their instructional practices for this group. Using this information, we find that the tutors provide much more individual instruction to low achievers compared to high achievers and that low achievers seem to spend quite a lot of time practicing basic skills, while high achievers are more likely to be offered problems that can be solved in several ways. The variation in instructional practices between the tutors – notably with respect to the degree of inclusive individualization for low achievers - is substantial.

In the second part of this paper, we investigate the associations between small group teachers' instructional practices and student performance by exploiting within-quintile variation in instructional practices across tutors. These analyses show that low achievers faced by tutors who practice inclusive instruction experience large achievement gains. While the performance of middle and high achievers is not associated with tutors who report that they practice intensive individualization for these subgroups, the performance of middle and high achievers is positively associated with tutors who report that they practice intensive individualization for low achievers - indicating that this subgroup of tutors has skills that benefit other subgroups of students as well.

Finally, we show that the treatment effects for schools with tutors who practice inclusive instruction for low achievers are almost twice the size of the treatment effects for schools without this type of tutors. The magnitude of the differences in treatment effects suggests that the instructional practices are correlated with other mediating factors. The degree of tailoring is one such factor. We provide indicative evidence that the homogeneity of the small groups is crucial for the size of the treatment effects.

The rest of the paper is organized as follows. In the two next sections we present details of the experiment, relevant institutional detail, and the data. The analyses have three parts. In the first part, we show how the tutors respond to the student body composition of the small groups by adjusting their instruction, and we show that the tutors can be classified into four types based on major characteristics of their instruction. In the second part, we show how the variation in achievement gains across treatment schools is associated with tutor types, and in the third part how the treatment effects vary across the four types of tutors. We discuss our contributions and conclude in the final section.

Background and characteristics of the field experiment

Background

In 2015, the Norwegian government decided to increase the teacher-to-student ratio in compulsory schooling. As part of this initiative, NOK .5 billion was allocated to research into the short-term effects of higher teacher-to-pupil ratios, and to two field experiments with a focus on pupils in early grades, one which experimented with reading instruction with two teachers in the classroom

the other experimenting with mathematics instruction using small homogenous groups. The present paper uses data from the latter experiment.

Institutional context and the intervention

The experiment had to be carried out within the framework of ordinary mathematics teaching in the public school (enrolling 96.3% of all students in 2016). The public schools are governed by a two-tier system. The national government sets goals, curriculum, distributes instructional time across subjects, defines minimum standards for teachers' formal qualifications and the maximum number of students per teacher. Inclusion is strongly emphasized. Thus, no student subgroups can be excluded from regular classrooms, except for shorter periods of time. The experiment, being a combination of in-school delivery and a pull-out strategy is adapted to these institutions.

First, it was accepted that six weeks are within the limit for a short period, so the treatment was decided to consist of two periods of small group intervention per school year, each period of 4-6 weeks in length. Second, the treatment dosage is determined by legislation saying that the students will be taught mathematics for 560 hours during grades 1-4, or on average 140 hours per year, implying that the treated students received instruction in small groups 30 to 44 hours per year. The sessions differed in length, as there are local variations in the schools' organization of the regular mathematics instruction. While some schools have long sessions (up to 90 minutes), others have shorter sessions, often 60 or 45 minutes, but always adding up to 140 hours per year. Instruction was given in parallel to all regular mathematics classes.

The local municipalities differ much in size, implying that the number of schools and students per municipality differs much, from one elementary school in the smallest municipalities to 107 schools in the capital Oslo in 2016. In 2016 the national average number of schools per municipality was 6.6.

Since the lion's share of field experiments with tutoring are carried out in the US, it should be noted that there is more between-school segregation by ability in the US than in Norway, where the variation in student performance is much larger within than between schools.

Randomization

10 large or quite large municipalities spread around Norway were invited to participate in the field experiment. Large municipalities were chosen because they have relatively well-functioning local labor market for teachers and reasonable staffed municipal administrations, implying that they might have the capacity to recruit the new teachers and keep control schools going for 4 years with taking tests and providing necessary information. In addition, this approach was chosen because it could shed some light on the local governing system as a moderator for treatment effects. The 10 superintendents were informed that participation would give half of the elementary schools in the municipality one extra teacher man-year (an average of 8 man-years per municipality).

We conducted stratified randomization in the following manner. Within each municipality the schools were ranked based on their mean test score in the national math tests at the fifth grade (no tests are taken at earlier stages). We averaged over the mean score in the two preceding school years (2014, 2015) to reduce measurement error. Next, we constructed a set of strata of

at least four schools in each stratum. In doing so, we followed the recommendation by Imbens (2011) to have at least two treatment and control schools in each stratum, so that one can derive a within-strata variance in the treatment effect. Most strata consist of four or six schools. In three municipalities, we had an uneven number of schools who volunteered to participate in the project, which resulted in one stratum in each municipality with seven schools. Next, we randomized schools to the treatment or the control group by using a random number generator. A total of 159 schools participated, 81 in the control group, 78 in the treatment group. Appendix Table 1, reproduced from our 2022-paper (Bonesrønning et al. (2022)), shows that randomization was successful.

Having informed municipalities and schools about the outcomes of the randomization process, the researchers visited all participating municipalities to present the intervention for municipal officers and school leaders in treatment and control schools. All schools were informed about the intervention. The leaders in control schools were told not to make changes in the use of resources, to participate in pre- and post-tests and to report on the school's organization of teaching. The treatment schools received information about the formation of small groups (size, composition, duration of small group treatment), about cooperation and coordination between the ordinary teachers and the small group teacher(s), and about the routines for reporting about small group participation. These meetings ensured that the information reached the schools widely, helped to clarify misunderstandings and mobilized the schools for implementation.

The implementation

Implementation was discussed with municipal officers and school principals in all the participating municipalities. Some compromises were made. The project leadership accepted that the school principals in treatment schools could decide whether to allocate the new teacher to small group instruction or substitute the new teacher for an existing staff member who then was allocated to small group instruction. A few schools asked to divide the teacher man-year into two parts. In this case, the two teachers should be responsible for the small group teaching in one of the two cohorts. Importantly, agreement was reached that there should be only one tutor per cohort in each school.

All schools – control schools as well as treatment schools - in the 10 municipalities were instructed to keep the number of teacher assistants in the intervention grades unchanged and not change the use of school resources due to the schools' participation in the project. Since in most schools the assigned teacher man-year was not fully filled up with small group teaching, the schools were instructed to use the rest of the man-year for grades that did not participate in the experiment.

In small schools (with one class per grade) or medium sized treatment schools (with two classes per grade) all students in the chosen grades were included. In schools with more than two classes in each grade one teacher man-year was not enough to provide treatment to all students. In these cases, the project leader randomized two classes to treatment. In our earlier intention-to-treat analyses (Bonesrønning et al, 2022) all classes in treatment schools with more than two classes were included as treated. In the present treatment-on-treated analyses only the treated classes are included.

The handbook, targeting participating teachers and containing much of the information from the introduction meetings, was distributed to all schools. Here the teachers were recommended

to form small groups that were homogenous with respect to pre-test scores to facilitate tailored instruction to students, and it was emphasized that the two teachers -in the regular class and the small group respectively- should cooperate to coordinate the teaching, to ensure seamless returns to the home class. Assessments should be used to guide areas for focus, provide feedback to students and track student progress. Connections should be made between out-of-classroom learning (in small groups) and classroom teaching.

Empirical evidence about the characteristics of effective instruction in mathematics, based on reviews of existing research made by What Works Clearinghouse (Gersten, Beckmann, Clarke, Foegen, Marsh, Star, and Witzel, 2009) and the National Mathematics Advisory Panel (2008) were presented in the Handbook.

Data and Research Design

Data

We use student-level data for two cohorts of students (the 2008- and 2009 cohorts) covering one year (2016/17) for the 2009-cohort and two years (2016/17 and 2017/18) for the 2008-cohort, a total of 16 276 students in the two cohorts. Appendix Table 2 provides information about the cohorts, treatment length, and pre- and post-tests. Privacy concerns dictate that survey data cannot be mixed with register data, implying that individual students can only be characterized by pre- and post-test results in the present study.

Frequent reporting to the project manager about the composition of the small groups was part of the job description for the tutors. That is, the tutors were asked to identify the students in the current small group, and the dosage of treatment measured by the number of weeks and the number of lessons per week so that the quantitative parts of the treatment could be described in detail.

All mathematics teachers involved in the experiment received questionnaires about their background (education and experience) and they were asked repeatedly about their instructional practices, especially about the allocation of available instructional time between presentations, seatwork, guided practice, and feedback, about their emphasis on automatization versus problem solving, and even more.

The students in treatment and control schools were tested in mathematics early in the fall of 2016 -a few weeks after start of the semester. Ideally, the tests should have been taken prior to treatment, but this could not be accomplished due to a strict timeline imposed on the project. The first post-test was given at the end of the first year of treatment. These tests were developed for the project by professionals familiar with test design and teaching in the early grades and piloted in schools outside the project. The tests were closely connected to the curricula for the respective grades. The tests were conducted by a company that specialized in testing, the tests were online, and the results were scored automatically.

Table 1 shows that the two cohorts have approximately equal sized small groups with an average of 4.9 students. The standard deviations are approximately 0.85, indicating that quite a few small groups exceed the upper limit of 6 students. The average dosage is 8.2 weeks for both cohorts, with standard deviations about 0.8, indicating that quite many students receive less than the minimum of 4x2 weeks of small group instruction per year. Obviously, even

though most schools are well within the limits set for size and dosage, some schools do not meet the minimum requirements for treatment. We consider the consequences of deviations from the requirements when discussing the robustness of the findings.

TABLE 1 *Descriptive statistics for the 2008- and 2009-cohorts. Small group size, dosage, and homogeneity.*

	2008-cohort		2009-cohort	
	Mean	St.Dev.	Mean	St.Dev.
Average number of weeks in each small group period	4.1	0.73	4.08	0.76
Average small group size	4.9	0.86	4.8	0.84
Total number of minutes in small group instruction	3031	973	2959	950
To what extent do you agree with the following statement: Students are placed into small groups with students on the same ability level (1-5 scale)	4.37	0.83	4.48	0.77

Note: Numbers refer to the treatment years 2016/17 and 2017/18 when both the 2008 and 2009 cohorts were treated. The 2009 cohort continued receiving treatment in the year 2018/19.

The last row in Table 1 shows that most tutors agree that the small groups are homogenous with respect to the pretest score. To describe the composition of the small groups more precisely we have ranked all students and all small groups by quintiles based on pretest scores. For the small groups, the rank is based on the average pretest score. If all schools were equal (the average pretest scores being equal to the sample mean and equal distributions), and if the students were perfectly sorted, the difference between group rank and individual rank would zero for all students. In Appendix Table 3 we show that 86-87 percent of the students belong to groups with ranks -1, 0 or 1.

80 teacher man-years are filled with teachers formally qualified to teach at the elementary level are hired by the schools. Observable characteristics of the tutors and the regular mathematics teachers are reported in Table 2.

TABLE 2 Characteristics of tutors and regular math teachers. Treatment schools

Teacher characteristics	Average	St.Dev.	Min	Max	N
Gender (female =1):					
Tutor	1,28	0,449	1	2	98
Regular teacher	1,13	0,337	1	2	195
Age:					
Tutor	40,1	11,23	24	66	94
Regular teacher	41,7	11,42	24	67	191
Experience:					
Tutor	11,1	9,11	0	36	99
Regular teacher	12,3	9,62	0	40	207
Credits:					
Tutor	58,0	37	0	240	94
Regular teacher	36,9	29,7	0	240	200
>2 yrs. math secondary school:					
Tutors	0,469	0,502	0	1	96
Regular	0,401	0,491	0	1	200

It is more likely that the tutors are men, and slightly younger and slightly less experienced compared to the mathematics teachers in the regular classes. The tutors have more credits in mathematics from the teachers' college and have taken more courses in mathematics in upper secondary school compared to the regular teachers. Note that the number of tutors exceed 80, reflecting that in some schools the tutor position is shared between two teachers. In these cases, the two tutors are assigned to different cohorts.

Research design

Results from the intention-to-treat (ITT) analyses are reported elsewhere (Bonesrønning et al, 2022). Results from the treatment-on-the-treated (TOT) analyses are reported in Appendix 2. The most important findings are that the treatment effects are larger in the TOT- than in the ITT-analyses, that all student subgroups benefit from treatment, that students in the 3rd quintile experience the largest treatment effects, that students in the 1st quintile experience larger treatment effects than students in the 5th quintile, and that the treatment effects are larger in schools with low average pretest scores than in schools with high average pretest scores.

Our main purpose is to investigate whether the tutors' instructional practices matter for the size and distribution of the treatment effects. In this respect, the most interesting findings from the

ITT- and TOT-analyses are that low achievers experience relatively large treatment effects, indicating that the tutors' instruction outbalance negative peer effects.

The analyses address the following research questions.

Research question 1: Do the tutors' instructional practices differ across student subgroups and schools?

Within cohorts-in-schools, the same tutor teaches all student subgroups. We can therefore address two long withstanding controversies in educational research: whether the teachers/tutors teach to the level of the students (individualization), and whether there is substantial between- teacher/tutor-variation in instruction for students at the same level.

Research question 2: Are the tutors' instructional practices associated with differential achievement gains across student subgroups?

Research question 3: To what extent can the tutors' instructional practices explain the observed variation in treatment effects?

Research questions 2 and 3 address variation in achievement gains within treatment schools and the difference in achievement gains across treatment and control schools, respectively.

The tutors' instructional practices

Existing evidence, model, and descriptive statistics

The existing empirical evidence on teachers' responses to the student body composition of classrooms is thin, scattered and speaks only indirectly to the current intervention. Tomlinson et al (2003) have reviewed the empirical literature on differentiation of instruction and conclude that there are indications "that most teachers make few proactive modifications based on learner variance." Morgan, Farkas and Maczuga (2015) have investigated the mathematics instructional practices of first-grade teachers to find that many teachers choose ineffective instruction, i.e., student-centered instruction for students with mathematics difficulties. More optimistic findings come from a tracking experiment in Kenya, where Duflo, Dupas, and Kremer (2011) provide indirect evidence that teachers respond to homogenous groups by tailoring their instruction to the students' skills - the effects of tailoring being large enough to dominate any negative peer effects for low achievers.

Betts and Shkolnik (1999) have investigated math teachers' response to reductions in class size to find that the "teachers shift time away from group instruction and towards individual instruction", but that "large reductions in class size shift teachers' time allocation by only a few percentage points." They also find that teachers react more strongly to class size changes when teaching below-average students.

We are aware of no empirical studies of how tutors allocate their instructional time across students in small groups. However, Guryan, Ludwig, Bhatt, Cook, Davis, Dodge, Farkas, Fryer, Mayer, Pollack, and Steinberg (2021) provide indicative evidence that "personalization" of instruction is a major mechanism behind the positive effects of a two-on-one math intervention for struggling 9th and 10th graders in Chicago public schools.

We interpret the existing empirical evidence as showing that the teachers instructional practices vary somewhat with the student body characteristics, but that this variation is limited by the teachers' preferences and skills. The analyses and discussions in this paper are structured according to the simple model presented in Figure 1.

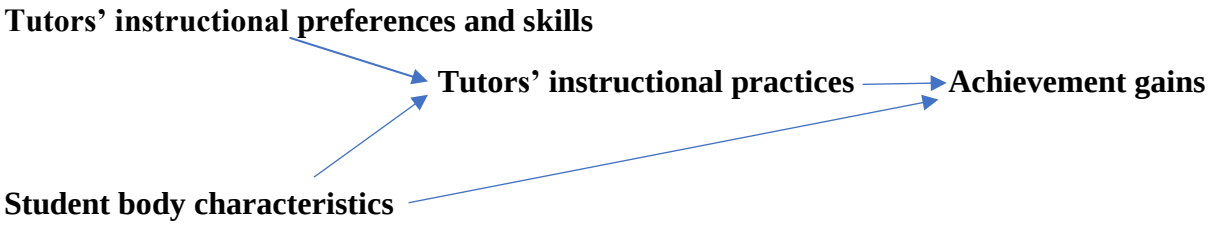

FIGURE 1 *Tutors' instructional practices. Determinants and effects*

The model distinguishes between direct and indirect peer group influences, where the latter is mediated through tailoring, i.e., the tutors' responses to the composition of the small groups. In addition, it is assumed that the variation in tutors' instructional practices across student subgroups reflect their underlying preferences and skills. This model is rich enough to explain the within-tutor variation in practices across student subgroups, and the between-tutor variation in instruction for the same type of student subgroups. That is, holding tutors' instructional preferences constant, variations in the tutors' instructional practices reflect variation in the student body composition. And vice versa, holding peer group characteristics constant, variations in the tutors' instructional practices reflect variation in the tutors' instructional preferences.

In our case, the within-school-across-student-subgroups variation is large, so we expect to find substantial variation in individual tutors' instruction across the small groups. By applying tutor fixed effects, we avoid the most serious obstacles for causal inference. That is what we are doing in this section.

To investigate the degree of tailoring, we separate between teacher-directed and student-centered practices. There is no consensus about how to operationalize these concepts for instruction in regular classrooms (see Morgan et al (2015) for a discussion). In the present intervention, the tutors face small groups of 4-6 students. In this situation, teacher-directed practices are essentially about the tutor reaching out with assistance to all the students in the group. Although the small groups are homogenous with respect to a pretest score, they are heterogenous in other respects - especially this goes for the low achievers' groups. The tutors' incentives to spend much time with the most struggling students and the evasive students are weak, so we expect that across-tutor variation in preferences and skills to be reflected in the allocation of instructional time within the low achievers' groups.

Student-centered practices are about the tutor organizing the students for cooperation. In this case, there is less competition for tutor's attention, but since cooperation makes no sense

without student involvement, we expect the tutors to mostly use these practices for high achievers. We add a content dimension by distinguishing between routine practice and drill on the one hand and working with problems that can be solved in different ways on the other.

Content may interact with the instructional practices: our hypotheses are that routine practice and drill increases the effects of individualized instruction, while providing problems that can be solved in different ways increase the effects of collaboration.

The data come from surveys where the tutors are asked to characterize the small group they currently are working with (four categories: low achievers, medium achieving students, high achievers, or mixed groups), and then to characterize their teaching for this group according to the teacher-directed vs. student-centered dichotomy. These examinations were taken four times during the second year of intervention. The information is used to establish within-tutor variation in instruction across student subgroups.

The individualization of instruction is measured by the tutors' responses to the following three statements: "I supervise students who need help" (indiv1), "I supervise individual students I know need help, even if they do not ask for help" (indiv2), and "I spend time with students who are not working unless I follow them up" (indiv3). As stated above, these statements are motivated by the often-reported observation that teachers give attention to the upper third in the class. Applied to instruction in small homogenous groups, we thus investigate whether - among students who perform at the approximately same level- the evasive and non-working students get least attention.

All measures are derived from rating of statements on a 1-5 scale where 1 is "strongly disagree" and 5 is "strongly agree". Table 3, the top panel, shows that the indiv1-measure has an average of 4.34 and a relatively small standard deviation of 0.54, indicating that most of the tutors say that they agree or strongly agree that they supervise students who need help. The proportion of tutors who practice intensive individualization, i.e., agree or strongly agree that they supervise students who do not ask for help, is much smaller, and the variation substantially higher, compared to the indiv1-measure. Even fewer tutors agree or strongly agree that they spend much time with non-working students (indiv3), probably reflecting that there are few such students present. The tutors' responses are reported separately for low achievers' and high achievers' small groups on the lower part of Table 3. Even though the numbers reported from low achieving groups are higher than those reported for the high achieving groups, high numbers are reported from the latter groups as well – indicating that these practices are widespread across the students' math skill distribution.

Student-centered practices are measured by the tutors' responses to the following statement: "The students solve problems together, the entire group" (studcent1). The average of 3.26 indicates that these practices are not widespread, and as can be seen, not even among tutors reporting from high achieving small groups. Information about the kinds of skills that are emphasized in the groups are generated from the following statements: "The students were given problems that can be solved in several ways" (Problems), and "We spend time automating arithmetic operations" (Automat). Neither of these practices seems to be much practiced, indicating that these statements do not tap the tutors' practices very well.

TABLE 3 *Tutors' instructional practices. Descriptive statistics. Teacher observation data (1-4 observations per teacher)*

Variable	Observations	Mean	St.Dev.
All students:			
Studcent1	264	3.25	0.97
Problems	260	3.55	0.91
Automat	259	2.84	1.01
Indiv1	266	4.34	0.59
Indiv2	266	3.80	0.93
Indiv3	260	3.17	1.25
Low achievers:			
Studcent1	69	3.26	0.98
Problems	67	3.30	0.97
Automat	66	3.15	1.04
Indiv1	68	4.44	0.58
Indiv2	68	3.91	0.94
Indiv3	67	3.54	0.97
High achievers:			
Studcent1	98	3.34	0.92
Problems	97	3.62	0.94
Automat	97	2.82	0.97
Indiv1	100	4.32	0.63
Indiv2	99	3.67	0.95
Indiv3	99	2.78	1.35

Note: 1-4 observations per tutor

Within-tutor variation in instruction across the small groups

The descriptive statistics presented in Table 3 are uninformative about the magnitude of the within-tutor variation in instructional practices across the small groups. We address this variation by estimating equations with tutors' instruction as dependent variables (6 elements) and small group composition - as reported by the tutors - as independent variables using a tutor fixed effects specification:

$$tutor\ instruction_{SG} = \beta_0 + \beta_1 LA_{SG} + \beta_2 MA_{SG} + \beta_3 HA_{SG} + \beta_4 \bar{y}_{st-1} + \mu_T + \varepsilon_s \quad (1)$$

where $LA_{SG}, MA_{SG}, HA_{SG}$ indicate that the small groups contain low achievers, medium achievers, or high achievers, respectively. The reference group is made up by small groups reported to have mixed compositions. μ_T is a tutor fixed effect. The results for the 6 estimated equations are reported in Table 4.

TABLE 4 *Tailoring of instruction. Fixed tutor effects*

Dependent variables:						
	Indiv1	Indiv2	Indiv3	Studcent1	Problems	Automat
Independent variables:						
Low achieving group	0.235* (0.131)	0.471** (0.203)	0.218 (0.288)	-0.154 (0.226)	-0.353* (0.195)	0.425 (0.259)
High achieving group	-0.221* (0.122)	-0.344** (0.169)	-0.729*** (0.246)	0.214 (0.186)	0.249 (0.182)	-0.087 (0.200)
Middle achieving group	-0.0105 (0.104)	0.177 (0.136)	0.0364 (0.223)	-0.145 (0.185)	0.190 (0.144)	-0.091 (0.184)
Tutor fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
N	268	268	262	266	261	262
R^2	0.567	0.576	0.534	0.579	0.547	0.539

Note: The reference group is made up of small groups with mixed compositions

Table 4 shows that the tutors tailor their instruction to the composition of the small groups, providing more individualized instruction to low achievers than to the other subgroups. The differences in instruction between groups of high- respectively low-performing groups are greater for the targeted indiv2- and indiv3-measures, than for the more general indiv1-measure.

High achievers collaborate significant more on problem solving than do low achievers. No student subgroups deviate significantly from the reference group with respect to the use of routine practices and drill, but these results obscure the fact that there is a statistically significant difference in the use of drill for low achievers relatively to high achievers.

The within-tutor variation in the indiv1-measure across small groups of low- and high-achievers is 0.45 points. The total variation in the indiv1-measure across the tutors as measured by the standard deviation is 0.59 points, implying that the tutors' response to the difference in small group composition between the extremes equals 0.76 SD (0.45/0.59) in the indiv1-measure. This indicates that even though the tutors' responses to the small group compositions are significant, they explain a relatively small fraction of the variation in the indiv1-measure. Put another way, the across-tutor variation in the indiv1-measure for low achievers is likely to be substantial.

The within-tutor variation in the indiv2- and indiv3-measures explains somewhat more of the total variation in these measures: the difference in small group composition between the

extremes equals 0.93 and 0.95 SD, respectively. This is as expected, since evasive and non-working students are likely to be less present in middle and high achievers' groups.

We have estimated an equation making use of the between-tutor variation in instructional practices as well, that is, the equation does not include tutor fixed effects but is otherwise like equation (1). The results are reported in Table 5 and show that the signs and statistical significance of the estimates coincide very much with the estimates in Table 4. However, the estimated differences in tutors' responses are substantially smaller in this case – providing evidence that the tutors' responses to the student body compositions are biased downwards when adding between-tutor variation in instructional practices. It should be noted that in this case there is more evident that the content differs across low and high achievers, as low achievers are exposed to significantly less problem solving and significantly more routine practice and drill.

TABLE 5 *Tailoring of instruction. Without fixed tutor effects*

Dependent variables:						
	Indiv1	Indiv2	Indiv3	Studcent1	Problems	Automat
Independent variables:						
Low achieving group	-0.150 (0.117)	0.00640 (0.197)	0.277 (0.233)	-0.210 (0.205)	-0.360* (0.203)	0.161 (0.230)
High achieving group	-0.342*** (0.111)	-0.351* (0.178)	-0.538*** (0.248)	0.0169 (0.167)	0.142 (0.172)	-0.304* (0.181)
Medium achieving group	-0.165* (0.0.092)	0.0440 (0.137)	0.223 (0.184)	-0.152 (0.151)	0.108 (0.130)	-0.295* (0.150)
Tutor fixed effects	No	No	No	No	No	No
N	268	268	262	266	261	262
R^2	0.037	0.026	0.055	0.008	0.031	0.031

Note: The reference group is made up of small groups with mixed compositions.

There are at least two take-aways from these analyses: First, the tutors vary their instruction across the small groups, emphasizing teacher-directed individualization for low achievers, and student-oriented problem solving for high achievers. Medium and high achievers spend less time on routine practices and drill than do low achievers. Second, the instructional practices vary substantially across the tutors.

Tutor types

We have characterized the tutors' instructional practices along three dimensions: teacher directed versus student centered, inclusive versus exclusive, and drill versus problem solving.

The correlation matrix in Table 6 shows that quite a few of the tutors who agree that they supervise students who need help – which we interpret as teacher-directed tutors - also agree

that they help students who do not ask for help – which we interpret as being inclusive tutors. Fewer of the tutors who spend much time supervising students who need help, spend much time on students who are not working unless followed up by the tutor.

TABLE 6 *Correlation matrix for measures of individualization*

	Indiv1	Indiv2	Indiv3
Indiv1	1		
Indiv2	0.4417	1	
Indiv3	0.2425	0.5089	1

Very few tutors report high values on indiv3. We therefore separate the tutors into four categories based on their responses to the indiv1- and indiv2-statements. The tutors in category HH report high values (4 or 5) for both statements, the tutors in category HL report a high value for indiv1 and a low value for indiv2, and so on. Tutors who report high values on the indiv1- and indiv2-measures spend more time with evasive students than do tutors in category HL, while tutors in the LH-category prioritize assistance to evasive students. We label the instructional practices of HH- and LH-tutors as inclusive individualization.

Table 7 shows how the tutors are distributed across the four categories. Three of the cells contain large numbers of observations. The tutor type LH is rare.

TABLE 7 *The distribution of tutors according to their instructional practices*

	Indiv1		
Indiv2		High	Low
	High	90	17
	Low	93	66

Note: The total number of observations is 266, reflecting that the tutors have responded 4 times

Most tutors are represented in several cells, implying that there is no unique classification of tutor types.

Associations between tutors' instruction and student achievement[2]

Here we address our second research question: Are the students' achievement gains associated with the tutors' instructional practices?

Because the classification of tutors is largely determined by the tutors' responses to evasive students, the association between achievement gains and the instructional practices can be mediated through a "third factor". Here we develop hypotheses about the relationships between individual students' achievement gains and the tutors' instructional practices by applying the assumption that unassisted evasive students experience small achievement gains. Our focus is

[2] We do not report associations between the tutors' instructional practices and their background characteristics because the instructional practices vary across student subgroups for fixed background characteristics.

on low achieving students. HH- and LH-tutors spend much time with evasive students, so we expect that these tutors are associated with large achievement gains for low achievers, independent on the proportion of evasive students within the subgroups of low achievers. LL- and HL-tutors report that they spend little time with evasive students. This could be because evasive students are absent in their small groups of low achievers or because such students are present but are not offered much attention. In both cases, unassisted low achievers are likely to experience small achievement gains for LL-tutors. We expect that the HL-tutors are associated with high achievement gains unless there are evasive students in their small groups. If evasive students are present, there could be large within-group differences in achievement gains in the HL-tutors' groups due to biased allocations of instructional time.

We investigate these hypotheses by estimating the following equation:

$$y_{ist} = \alpha_0 + \alpha_1 y_{is,t-1} + tutor\ instruction_s \alpha_2 + \alpha_3 \bar{y}_{st-1} + \varepsilon_{ist} \qquad (2)$$

where y_{ist} and $y_{is,t-1}$ are post- and pretest scores for student *i* in school *s*, respectively. $\bar{y}_{st-1}$ is the average pretest score for the peers in the small group, and $tutor\ instruction_s$ is the vector describing the tutors' instructional practices, including the tutor types from Table 7, in addition to the extent of group work, problem solving, and automatization. We estimate the equation separately for low achievers, medium achievers, and high achievers, making use of the subgroup-specific measures of tutors' instruction. The results are presented in Table 8.

TABLE 8 *Associations between student achievement and tutors' instruction – 4 tutor types*

	Low achievers	Medium achievers	High achievers
Pretest, individual	0.643***	0.411***	0.742***
	(0.0457)	(0.132)	(0.0502)
Pretest, average	-0.353***	-0.342***	-0.341***
	(0.0986)	(0.0838)	(0.0726)
Studcent1	0.0641	-0.0111	0.0003
	(0.0433)	(0.0389)	(0.0295)
Problems	-0.0159	-0.0707*	0.0234
	(0.0358)	(0.0402)	(0.0316)
Automat	-0.107**	0.0301	-0.0581*
	(0.0419)	(0.0402)	(0.0299)
HH	-0.0911	-0.0408	0.135*
	(0.0916)	(0.0818)	(0.0811)
HL	-0.671***	0.0826	-0.107
	(0.192)	(0.287)	(0.129)
LL	-0.255*	0.0397	0.00694
	(0.129)	(0.102)	(0.0729)
Constant	0.360	0.350*	0.0741
	(0.220)	(0.314)	(0.156)
N	1283	1093	1343
R^2	0.213	0.044	0.132

Notes: Dependent variable is standardized individual posttest score. Tutor type LH is the reference category for the individualization variables. Robust standard errors in parentheses. ***p<0.001, **p<0.05, *p<0.1

Table 8 shows that the tutors who report high values on the indiv2-measure (HH, LH) are associated with significantly larger achievement gains for low achievers compared with tutors who report low values on this measure (HL, LL). Low achievers exposed to tutors who report low values on the indiv2-measure gain on average 0.3 to 0.7 SD less than the students exposed to LH-tutors. These findings are consistent with the hypothesis that tutors who assist evasive students respond properly to the challenges within the low achievers' groups. Low achievers exposed to HL-tutors do poorly, which is consistent with the hypothesis that evasive students are present but not offered much attention from these tutors. Table 8 also shows that high achievers exposed to HH-tutors perform significantly better (at the 0.1 significance level) than high achievers exposed to the other tutor-types, LH-tutors included. These findings might indicate that LH-tutors have a stronger inclination to prioritize struggling students and are more withdrawn compared to HH-tutors in high achievers' small groups.

The estimates for the tutors' instructional practices differ much for low achievers but not for middle and high achievers. One potential explanation for this difference is that our characterization of the instructional practices is irrelevant for subgroups others than low achievers. Inspired by existing empirical research on teacher effectiveness (for example Chetty, Friedman, and Rockoff, 2014) which show that teachers who succeed with one class often succeed with other classes, we investigate whether the tutors who succeed with low achievers also succeed with middle and high achievers. Equation (2) is estimated for medium and high achievers using the classifications of tutors as reported for low achievers (HH_LA, LH_LA….). The results from this exercise are presented in Table 9.

TABLE 9 *Associations between student achievement and tutor type as defined by their response to low achievers*

	Low achievers	Medium achievers	High achievers
Pretest, individual	0.643***	0.414***	0.740***
	(0.0457)	(0.152)	(0.055)
Pretest, average	-0.353***	-0.260***	-0.286***
	(0.0986)	(0.0864)	(0.0832)
Studcent1	0.0641	-0.0725	0.0467
	(0.0433)	(0.0492)	(0.0370)
Problems	-0.0159	-0.0246	0.0554*
	(0.0358)	(0.0416)	(0.0330)
Automat	-0.107**	-0.0400	-0.0269
	(0.0419)	(0.0427)	(0.0315)
HH_LA	-0.0911	0.0233	0.0910
	(0.0916)	(0.0961)	(0.0805)
HL_LA	-0.671***	-0.582*	-0.249*
	(0.192)	(0.296)	(0.144)
LL_LA	-0.255*	0.00715	0.0654
	(0.129)	(0.102)	(0.0838)
Constant	0.360	0.160	-0.247
	(0.220)	(0.184)	(0.193)
N	1283	602	1151
R^2	0.213	0.048	0.130

Notes: Dependent variable is standardized individual post test score. Tutor type LH_LA is the reference category for the individualization variables. Robust standard errors in parentheses. ***p<0.001, **p<0.05, *p<0.1

Table 9 shows that middle and high achievers who are exposed to HL_LA-tutors perform significantly poorer than the middle and high achievers who are exposed to LH_LA-tutors - the estimate for middle achievers being more than twice the size of the estimate for high achievers. These results reflect that the subgroups of HL-tutors as identified from their responses to low achievers differ in important respects from those tutors identified as being of the HL-type from their responses to middle and high achievers. Tutors who report that evasive low achievers receive little attention, appears to do poorly with the other student subgroups as well – but we cannot tell why they do poorly.

Can the tutors' instructional practices explain the treatment effects?

The results reported in Tables 8 and 9 are based on variation in achievement gains across treated schools. Here we investigate how the treatment effects vary with the tutors' instructional practices. We estimate treatment-on-treated effects (equation A1) by quintiles and for subcategories of schools based on the tutors' instructional practices. The treatment schools are divided into two subgroups of inclusive (HH and LH) and non-inclusive tutors (LL and HL) – where the classification of tutors is based on their instructional practices for low achievers. The discussion of the relationships between tutor instruction and the student body composition prior to the presentation of equation (2) is still relevant. These challenges are more serious for low achievers than for middle and high achievers. The results from estimation of treatment-on-treated equations are presented in Table 10.

TABLE 10 *Treatment effects by quintiles for schools with inclusive and non-inclusive individualizing tutors*

	All	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Schools with inclusive individualizing tutors						
Treatment	0.228***	0.257***	0.233***	0.240***	0.214***	0.198***
	(0.0356)	(0.0654)	(0.0507)	(0.0509)	(0.0448)	(0.0420)
Schools with non-inclusive individualizing tutors						
Treatment	0.126***	0.100*	0.0920***	0.199***	0.139***	0.0916**
	(0.0326)	(0.0564)	(0.0499)	(0.0523)	(0.0411)	(0.0404)

Note: The dependent variable is standardized individual posttest score. Independent variables in addition to the treatment indicator are standardized pretest score, mean pretest score, regular class size, and cohort. ***p<0.001, **p<0.05, *p<0.1

Table 10 shows that students in schools with inclusive individualizing tutors experience an average treatment effect of 0.23SD, with little variation across the quintiles. Students in schools with no-inclusive individualizing tutors (LL- and HL-tutors) experience an average treatment effect equal to 0.13SD, with larger across-quintile variation. Notably, the low achievers in quintiles 1 and 2 experience small treatment effects when exposed to LL- and HL-tutors. Thus, students in the 1st quintile experience the largest treatment differences (0.16 SD) across the two

types of tutors, closely followed by students in the 2nd quintile. The latter findings are consistent with the hypothesis that low achievers benefit much from inclusive and teacher-directed instruction.

For middle achievers – represented by students in the 3rd quintile – the treatment effects are relatively high and do not vary much across tutor types. For students in the 5th quintile, the results largely show the opposite: the treatment effects are smaller and the students who are exposed to HH- and LH-tutors experience more than twice the treatment effect experienced by students with student-centered tutors.

Looking across Tables 9 and 10, the results coincide to the extent that both approaches show that low achievers benefit greatly from being exposed to HH- and LH-tutors. In Table 10 it is much more evident than in Table 9 that students in the 5th quintile benefit from being exposed to HH- and LH- tutors compared to LL- and HL-tutors.

However, the relatively small treatment effects for high achievers, also when faced by HH- and LH-tutors, is contrary to expectations as conventional wisdom says that high achievers are likely to be the student subgroup that benefit most from tracking. We shed some light on the puzzle by estimating treatment effects separately for the two subgroups of effective tutors. Our hypothesis is that LH-tutors are more oriented towards evasive students and thus relatively more effective for low achievers. Table 11 reports the results.

TABLE 11 *Treatment effects for schools with different types of intensive individualization tutors*

	All	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Schools with HH-tutors						
Treatment	0.238***	0.217**	0.196**	0.248***	0.291***	0.233***
	(0.0613)	(0.110)	(0.0843)	(0.0792)	(0.0562)	(0.0756)
Schools with LH-tutors						
Treatment	0.212***	0.268***	0.232***	0.228***	0.170***	0.168***
	(0.0400)	(0.0795)	(0.0584)	(0.0560)	(0.0588)	(0.0481)

Note: The dependent variable is standardized individual posttest score. Independent variables in addition to the treatment indicator are standardized pretest score, mean pretest score, class size, and cohort. ***p<0.001, **p<0.05, *p<0.1

As shown, LH-tutors are associated with large treatment effects for students in the 1st and 2nd quintiles, larger than the comparable treatment effects for HH-tutors. HH-tutors are associated with larger treatment effects for students in the 4th and 5th quintiles - the difference in treatment effects between the two subgroups of tutors equal to 0.12 SD and 0.07 SD for students in the 4th and 5th quintiles, respectively. A suggestive interpretation is that the reported allocations of instructional time for low achievers reflect the tutors' preferences.

Discussion

Findings about mechanisms. Our first finding is that treatment varies across student subgroups within schools as low achievers are offered more individualized instruction than are medium and high achievers. Also, treatment varies substantially across schools, as the tutors provide more assistance and spend much time with evasive students in some of the schools.

Our second finding is that low achievers experience significantly larger *achievement gains* in schools where the tutors practice inclusive individualization. This subgroup of tutors is associated with the largest achievement gains for medium and high achievers as well.

Our third finding is that *the treatment effects* are almost twice as great in schools where the tutors practice inclusive individualization for low achievers compared to schools where the tutors report to spend little time with evasive students. The large difference in treatment effects may indicate that inclusive individualization correlates with other and unobservable tutor characteristics.

Our fourth finding is that the (large) treatment effects for medium achievers – notably students in the 3rd quintile - seem to be rather robust for variations in tutors' instructional practices, indicating that there are other mechanisms present.

The relatively large differences in treatment effects between subgroups of tutors motivate the question of whether there are additional mechanisms that are correlated with the instructional practices. We could think of several. The first is about the assignment of students to the small groups. Although the student body compositions of the small groups are largely determined by the composition of the regular classes, the tutors are likely to influence the distribution of students near the boundaries of the groups. Being the small fish versus the large fish in a group might have consequences for the treatment effects, and the tutors might hold different opinions about where individual students can flourish. The second is that the cooperation between the tutors and the regular class teachers might vary with tutor characteristics. We will return to these questions in future analyses.

Contributions to the tutoring literature. Unlike most existing field experiments with tutoring, we have access to black box information. By using this information, we find that the challenges related to biased allocations of instructional time across students seem to be persistent even in small homogenous groups – at least for a quite large subgroup of tutors. Also, we provide indicative evidence that tutors who spend little instructional time with evasive students are associated with small treatment effects. These consequences of biased resource allocations echo much existing empirical literature about struggling students being poorly treated in regular classes.

The importance of instructional quality is well recognized in existing experiments for struggling students, as tutors are often trained in advance to perform standardized instruction. Gersten et al (2015) implement the program Number Rockets and attribute the large treatment effect of 0.34 SD to a combination of pre-training of the tutors and a strict plan covering the topics, frequency of assessments and cumulative reviews. Guryan et al (2021) implement a program developed by Saga Education for 9th and 10th graders struggling with mathematics and report an average treatment effect of 0.37 SD. In their study, tutors were qualified for the job by participating in approximately 100 hours of training prior to the start of the school year. Like Number Rockets, the tutoring procedures - reflecting current knowledge about effective

tutoring - were described in detail. In addition, each school was overseen by a site director who "handled behavioral issues in the tutoring room and offered daily feedback and professional development."

One (indeed speculative) way to read the differences between the US and Norwegian experiments is that the US interventions report larger treatment effects because they provide guidelines and training that prevent quality variation across tutors, and on the other hand make use of small groups with 2-3 students that allow the tutors to spend much instructional time with each student. While Gersten et al. (op.cit) do not address potential mechanism, Guryan et al. (op.cit.) provide indirect evidence that personalization of instruction might be an important mechanism. At the same time, they are puzzled by the "enormous differences" in the estimated average effects between the two quartiles of students who had the most, respectively the least, benefits from the treatment. The latter findings echo our findings that the treatment effects differ within the subgroups of low achievers (the 1st and 2nd quintiles). A hypothesis for future investigations is that biased allocations of instructional time - even within very small groups - is a problem that is hard to avoid.

A unique feature of the Norwegian experiment is that middle and high achievers are offered tutoring. We have shown that the tutors respond to the average pretest scores in the small groups by adjusting their instruction from being teacher-directed to being more student-centered as they move from low achievers to high achievers. However, we have not addressed the importance of tailoring of instruction versus the importance of individualization of instruction, that is, in the present paper we cannot tell whether the size of the small groups is more important than the homogeneity of the small groups. The findings by Duflo et al (2012) from a tailoring experiment in Kenya - showing that tailoring is more effective than no-tailoring – indicate that group homogeneity is important for the treatment effects.

Limitations. The Norwegian field experiment is designed to test whether small group instruction leads to better student achievement, not to investigate how specific instructional practices affect students across the ability distribution. Since our purpose in the present paper is to examine the importance of the tutors' instructional practices for the size of the treatment effects, causality is very much out of reach. We report associations between student achievement gains and the tutors' instructional practices, and state that large differences in treatment effects across tutors who differ in their instructional practices, most likely reflect additional factors that are correlated with the tutors' instructional practices. Our approach to problems related to omitted variables is to capitalize on existing empirical evidence. Long existing evidence (Walberg, 1984, Bloom, 1984) indicate that tutorial instruction is by far the most important variable in the education production function, and therefore a sensible starting point. Our expectation is that results for the instructional practices as reported here, will be modified in our subsequent analyses which highlight issues such as peer influences and the quality of the collaboration between the tutor and the regular teacher.

Another potential concern is that the tutors' subjective assessments of the statements about characteristics of their instructional practices are not sufficiently reliable. At this point, we share the judgements made by Morgan et al. (2015): "self-report ratings are known to provide fairly accurate estimates of a teacher's relative frequency of use of particular instructional practices, and to co-vary with direct observation."

Moreover, the statements presented to the tutors -although inspired by the teacher-directed versus student-oriented dichotomy - are basically designed to examine the extent to which the tutors take advantages of the small groups to reach out to all students. Important features of the tutors' instructional practices are left out. Some of these omitted variables - such as the use of frequent corrective feedback - we should not worry about. Following Bloom (op.cit), who states that "the need for corrective work under tutoring is very small", we think this is a second-order worry. Other omissions are more problematic: by applying richer descriptions of the tutors' instructional practices, more could have been revealed about the properties of effective instruction for medium and high achievers.

Policy implications. The main motivation behind the intervention is to find ways to improve the mathematical performance of young students. Should we recommend local governments to spend additional money on small group instruction?

First, tutoring implies large costs which must be balanced by sufficient large treatment effects. In our previous paper (Bonesrønning et al., 2022) we have shown that the average effect is at least 0.12 SD per 1000 dollar which should be compared to 0.08 SD per 1000 dollar in STAR (Schanzenback, 2006). While the benefit-cost ratio is reasonable high, the contribution of the present analyses is to show that the benefits from small group instruction depend crucially on the quality of the tutors' instruction.

On one hand, we have shown that the variation in tutor quality is substantial. We cannot say whether quality differences reflect innate skills, or whether tutors can learn how to best utilize small and homogeneous groups to the advantage of the students. If it is about innate skills, scaling-up can be hard, especially in rural areas with thin teacher labor markets. If effective tutoring can be learned, we do not yet know whether this is achieved in advance or through on-the-job training. Studies in the US indicate that tutors learn in both ways – at least when it comes to manage very small groups for struggling students.

On the other hand, we have shown that a relatively large proportion of the tutors in the present intervention is associated with large treatment effects. A reasonable hypothesis is that the intervention has been attractive to highly skilled mathematics tutors. If small group instruction is only offered to a subgroup of schools with poor pretest scores and only to those students who struggle in mathematics, it is not unlikely that high-quality tutors would be less attracted to the intervention.

In scaling up, the government could follow one out of two models. The "Norwegian model" might be appropriate in areas with reasonable supply of high-quality mathematics tutors, and where improving the performance of all student subgroups is a high priority. The "US model" is more appropriate when the performance of struggling students is the number-one priority. A major advantage of the US model is that the tutors' tasks are much simpler, implying that they can be recruited from a much larger pool of applicants and can be qualified in advance through a relatively short period of training.

References

Betts, J.R., (2011). "The Economics of Tracking in Education", in Hanushek, Eric A., S. Machin, S. & Woessmann, L. (Eds.), *Handbook of the Economics of Education*, Volume 3, Amsterdam: North Holland, 341-381

Betts, J.R., & Jamie L. Shkolnik, J.L. (1999) The Behavioral Effects of Variations in Class Size: The Case of Math Teachers. *Educational Evaluation and Policy Analysis,* Vol. 21(2), 193-213

Bloom, B. S. (1984). The 2-sigma problem: The search for methods of group instruction as effective as one-on-one tutoring. *Educational Researcher,* 13(6), 4–16

Bonesrønning, H., Finseraas, H., Hardoy, I., Iversen, J.M.V., Nyhus, O.H., Opheim, V., Salvanes, K.V., Sandsør, A.M.J., & Schøne, P. (2021). Small Group Instruction to Improve Student Performance in Mathematics in Early Grades: Results from a Randomized Field Experiment CESifo Working Paper No. 9443, Revise & Resubmit, *Journal of Public Economics*

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review*, 104 (9): 2593-2632.

Clements, D. H., & Battista, M. T. (1990). Constructivists learning and teaching. *Arithmetic Teacher,* 38, 34–35.

Dietrichson, J., Bøg, M., Filges, T., & Jørgensen, A-M.K., (2017). Academic Interventions for Elementary and Middle School Students With Low Socioeconomic Status: A Systematic Review and Meta-Analysis. *Review of Educational Research,* Vol 87(2), 243-282

Dobbie, Will, and Roland G. Fryer Jr. 2013. "Getting beneath the Veil of Effective Schools: Evidence from New York City." *American Economic Journal: Applied Economics*, 5 (4): 28-60.

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *American Economic Review,* 101 (5), 1739-74

Gersten, R., Beckmann, S., Clarke, B., Foegen, A., Marsh, L., Star, J.R., & Witzel, B. (2009). Assisting Students Struggling with Mathematics: Response to Intervention (RtI) for Elementary and Middle Schools. Institute of Education Sciences (ED), National Center for Education Evaluation and Regional Assistance; What Works Clearinghouse (ED)

Gersten, R., Rolfhus, E., Clarke, B., Decker, L. E., Wilkins, C., & Dimino, J. (2015). Intervention for first graders with limited number knowledge: Large-scale replication of a randomized controlled trial. *American Educational Research Journal*, *52*(3), 516-546

Guryan, J., Ludwig, J., Bhatt, M.P., Cook, P.J., Davis, J.M.V., Dodge, K., Farkas, G., Fryer Jr., R.G., Mayer, S., Pollack, H., & Steinberg, L. (2021). Not Too Late: Improving Academic Outcomes Among Adolescents Jonathan Guryan, Jens Ludwig, Monica P. Bhatt, Philip J. Cook, Jonathan M.V. Davis, Kenneth Dodge, George Farkas, Roland G. Fryer Jr, Susan Mayer, Harold Pollack, and Laurence Steinberg NBER Working Paper No. 28531

Imbens, G. (2011). Experimental Design for Unit and Cluster Randomized Trials. International Initiative for Impact Evaluation Paper.

Morgan, P.L., Farkas, G., & Maczuga, S. (2015). Which Instructional Practices Most Help First-Grade Students With and Without Mathematics Difficulties*? Educational Evaluation and Policy Analysis*, Vol. 37(2), 184–205

Nickow, A., Oreopoulos, P., & Quan, V. (2020*).* The Impressive Effects of Tutoring of PreK12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence. NBER Working Paper No. 27476

Schanzenbach, D.W (2006) What Have Researchers Learned from Project STAR? *Brookings Papers on Education Policy,* No.9, 205-228

Stein, M., Silbert, J., & Carnine, D. W. (2004). Designing effective mathematics instruction. (4th ed.). Upper Saddle River, NJ: Prentice Hall.

The National Mathematics Advisory Panel (2008). Foundations for Success: The Final Report of the National Mathematics Advisory Panel. US Department of Education

Tomlinson, C.A., Brighton, C., Hertberg, H., Callahan, C.M., Moon, T.R., Brimijoin, K., Conover, L.A. & Reynolds, T. (2003). Differentiating Instruction in Response to Student Readiness, Interest, and Learning Profile in Academically Diverse Classrooms: A Review of Literature. *Journal for the Education of the Gifted*, Vol. 27, No. 2/3, 119–145

Walberg, H. J. (1984) Improving the Productivity of America's Schools. *Educational Leadership*, Vol.41, No.8, 19-27

APPENDIX TABLE 1 *Balance test*

	Control		Treatment		Difference
	N/[Schools]	Mean/SE	N/[Schools]	Mean/SE	(1)-(2)
Female	8128	0.481	8148	0.488	-0.007
	[81]	(0.006)	[78]	(0.007)	
Parental edu: Primary	8128	0.055	8148	0.054	0.001
	[81]	(0.007)	[78]	(0.007)	
Parental edu: Secondary	8128	0.213	8148	0.196	0.017
	[81]	(0.012)	[78]	(0.013)	
Parental edu: College, low	8128	0.390	8148	0.373	0.017
	[81]	(0.009)	[78]	(0.009)	
Parental edu: College, high	8128	0.308	8148	0.339	-0.031*
	[81]	(0.019)	[78]	(0.019)	
Parental edu: Missing	8128	0.035	8148	0.039	-0.004
	[81]	(0.003)	[78]	(0.004)	
Foreign-born	8128	0.063	8148	0.064	-0.000
	[81]	(0.005)	[78]	(0.004)	
Second generation	8128	0.100	8148	0.101	-0.002
	[81]	(0.011)	[78]	(0.013)	
School size	8128	56.615	8148	58.579	-1.964
	[81]	(2.153)	[78]	(2.238)	
F-stat joint significance, p-value					1.04, .41

Notes: Standard errors are clustered at school. Strata and cohort FE are included in all estimations. *** p<0.01, ** p<0.05, * p<0.1

APPENDIX TABLE 2 *Starting age and treatment duration*

	Cohort			
School year	2008	2009	2010	2011
2016/17	3rd grade$^{PRE, POST}$	2nd grade$^{PRE, POST}$		
2017/18	4th grade	3rd grade^{POST}		
2018/19	Test (5th grade)	4th grade		2nd grade$^{PRE*, POST}$
2019/20		Test (5th grade)	4th grade^{PRE*}	3rd grade
2020/21			Test (5th grade)	
2021/22				Test (5th grade)

Notes: The table shows the treatment age and duration of the four cohorts that were part of the 1+1 project as well as the timing of the different mathematics tests. PRE refers to the pre-test (baseline), POST refers to post-tests after treatment and Test refers to the National test for all 5th graders in Norway. *Completed in the spring at the end of the previous school year.

APPENDIX TABLE 3 *The distribution of individual students' rank*

Rank	Numbers	2008-cohort	2009-cohort
-4	5	0	5
-3	65	21	44
-2	292	106	186
-1	1,190	499	691
0	2,827	1,397	1,430
1	1,238	700	538
2	389	207	182
3	58	33	25
4	6	4	2

APPENDIX 1 ***Treatment-on-the-treated effects for quintiles of students.***

Bonesrønning et al (2022) report medium-term average treatment effects and some types of heterogeneity in treatment from ITT-analyses based on the same two cohorts as used in the present paper. Below we report short-term treatment effects from the TOT-analyses, emphasizing treatment effects for quintiles of students – which is relevant background information for the analyses of tutors' instructional practices.

The equations

The intent-to-treat (ITT) analyses show that the average treatment effects are positive and statistically significant. Those analyses make use of post-results from the 5th grade national test, a test taken 4 months after end of treatment. We make use of scores post-tests taken immediately after treatment is finished (for the year). These tests are more closely linked to the mathematics curricula than the national tests. All project-specific tests are piloted before use.

We start by estimating a standard treatment-on-treated (TOT) equation by quintiles based on the students' baseline test results. The estimated equations are:

$$y_{ist} = \beta_0 + \beta_1 y_{is,t-1} + \beta_2 \bar{y}_{-is,t-1} + \beta_3 T + \vartheta_c + \mu_{str} + \varepsilon_{ist} \qquad (A1)$$

where y_{ist} and $y_{is,t-1}$ are the post-test and pre-test scores for student *i* in school *s*, $\bar{y}_{-is,t-1}$ is the average pretest score in the grade, and T is the treatment indicator, ϑ_c is a cohort dummy and μ_{str} is a fixed strata effect. The students included are those that are present in the small groups at the pre- and posttests.

Treatment effects

Appendix Table 4 presents the results from estimation of equation (A1). The treatment effects are positive and precisely estimated for all quintiles, varying from 0.15σ to 0.23σ. Students in

the 3rd and 5th quintiles experience the largest and the smallest treatment effects, respectively - implying the lowest achievers benefit more from treatment than the highest achievers. These results are robust to alternative specifications: the estimates change slightly when estimating the equation without the two pretest variables. The quintile differences in effects are unaffected by the inclusion of class size among the independent variables.

APPENDIX TABLE 4 *Treatment effects across quintiles of students*

	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Treatment	0.198***	0.176***	0.231***	0.178***	0.145***
	(0.0477)	(0.0408)	(0.0375)	(0.0322)	(0.0303)
Average pretest	-0.319***	-0.305***	-0.233***	-0.268***	-0.139***
	(0.0737)	(0.0634)	(0.0548)	(0.0504)	(0.0421)
Individual pretest	0.493***	0.625***	0.681***	0.685***	0.871***
	(0.0504)	(0.0688)	(0.0909)	(0.0862)	(0.0668)
Constant	-0.451***	-0.127	-0.0211	-0.0383	-0.249*
	(0.131)	(0.0927)	(0.104)	(0.110)	(0.118)
Observations	2495	2749	2820	2797	2591
R^2	0.124	0.086	0.097	0.083	0.118

Notes: Dependent variable is standardized individual post test score. Robust standard errors in parentheses. ***p<0.001, **p<0.05, *p<0.1

Appendix Table 4 shows that the posttests - for fixed individual pretest scores - are lower in schools with high average pretest scores. In Appendix Table 5, where we have estimated equation (1) by quintiles separately for schools with high and low average pretest scores, respectively, we show that the largest treatment effects occur in schools characterized by low average pretest scores. The largest differences in treatment effects occur for students in the 1st quintile: these students experience almost twice as great treatment effects in low-performance schools as in high-performance schools. Note however that the treatment effect for low achievers in schools with high average pretest score is imprecisely estimated due to a relatively small number of observations. Students in quintile 3 stand out by experiencing the largest treatment effects in both subgroups of schools, and by deviating substantially from other student subgroups in schools with high average pretest scores.

APPENDIX TABLE 5 *Across-quintile variation in treatment effects across schools with different average pretest scores*

	1.quintile	2.quintile	3.quintile	4.quintile	5.quintile
Treatment- low average pretest score	0.246***	0.222***	0.283***	0.260***	0.230***
	(0.0597)	(0.0591)	(0.0632)	(0.0510)	(0.0551)
Treatment- high average pretest score	0.127	0.147**	0.246***	0.147***	0.124***
	(0.0771)	(0.0587)	(0.0508)	(0.0412)	(0.0378)

Notes: Dependent variable is standardized individual post test score. Independent variables are as in Appendix Table 2. Robust standard errors in parentheses. ***p<0.001, **p<0.05, *p<0.1